

DOCUMENT RESUME

ED 414 329

TM 027 857

AUTHOR Melnick, Steven A.; Henk, William A.
 TITLE Content Validation: A Comparison of Methodologies.
 PUB DATE 1997-02-21
 NOTE 12p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Hilton Head, SC, February 1997).
 PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
 EDRS PRICE MF01/PC01 Plus Postage.
 DESCRIPTORS *Classification; Comparative Analysis; *Content Validity; *Graduate Students; Higher Education; Reading Instruction; *Research Methodology; Self Evaluation (Individuals); *Test Content; Test Reliability
 IDENTIFIERS *Forced Choice Judgmental Review; *Latent Category Judgmental Review

ABSTRACT

This paper compares two methods of establishing content validity, forced-choice judgmental review and a latent category judgmental review. It also compares content validity evidence with the results of a scale reliability analysis and makes recommendations of the two content validity procedures. Two different groups of graduate students enrolled in a graduate program for reading specialists acted as expert reviewers for the content validation stage of the Reader Self Perception Scale (RSPS). Thirty students reviewed the items using the forced choice method of Gable and Wolf (1993) and the other 33 reviewed items using a latent category judgmental review process modified from that of Wiley (1967). In addition, the RSPS was administered to 2,733 fourth, fifth and sixth graders. While all test items were placed in the anticipated a priori categories by the forced choice reviewers, latent category reviewers identified finer distinctions among the items. It may be that the latent category method provides more accurate information with more distinctions among latent constructs. Reliability analysis of RSPS responses suggests that all items intercorrelate sufficiently and contribute to overall scale reliability. (Contains five tables and five references.) (SLD)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

Tm

ED 414 329

Content Validation: A Comparison of Methodologies

Steven A. Melnick and William A. Henk

Pennsylvania State University at Harrisburg

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

Steven Melnick

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to
improve reproduction quality.
- Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

TMO27857

Paper presented at the annual meeting of the Eastern Educational Research Association
Hilton Head, SC, February 21, 1997

BEST COPY AVAILABLE

Content Validation: A Comparison of Methodologies

Steven A. Melnick and William A. Henk

Pennsylvania State University at Harrisburg

Objectives

According to Gable and Wolf (1993), content validation should receive the highest priority during the process of instrument development. Unfortunately, many researchers, particularly the growing number of action researchers (i.e., teachers-as-researchers) do not appreciate its importance and consequently give scant attention to this crucial process. This lack of attention is due in part to unfamiliarity with the importance of content validity in addition to an uncertainty regarding the procedures. The purpose of this paper is to (1) compare two methods of establishing content validity (forced-choice judgmental review and a latent category judgmental review), (2) compare the content validity evidence with the results of a scale reliability analysis, and (3) to make recommendations regarding the two content validity procedures.

Theoretical framework

Content validity evidence is typically judgmental and can be obtained in different ways. A number of researchers (e.g., Delcourt & Kinzie, 1993; Gable & Wolf, 1993; Swanson, Tokar & Davis, 1994) recommend or utilize a judgmental procedure in which reviewers are first provided with concise descriptions (conceptual definitions) of each proposed category represented on the instrument. Typically, each category (i.e., construct) the instrument purports to measure is clearly defined and labeled. Reviewers are then asked to read each item carefully and indicate which of the

proposed categories it best "fits." In addition, reviewers are asked to indicate how strongly they feel the item fits the category. The data are analyzed by computing frequency of response percentages for each item by category. Gable and Wolf recommend a criterion level of 90% for an item to remain in that category without revision. Assuming that items receive at least 90% agreement in the *a priori* category the developer intended provides evidence of content validity. Items not meeting this criterion are either modified or deleted. One common criticism of this method is that the developer is "driving" the process by specifying the exact number of categories to which a reviewer can assign an item. In so doing, other potential distinctions a reviewer might "see" are lost.

A second, more empirical, method is called latent partition analysis (Wiley, 1967). In this procedure, reviewers are given a deck of cards with one item on each card. Reviewers are asked to read all items carefully and to sort the items into as many "meaningful and mutually exclusive" categories as they deem appropriate. These data are then analyzed statistically to determine if there are underlying meaningful content categories that reflect the judges ordering of the items. The strength of this approach is that the judgmentally derived categories can be compared to the *a priori* categories specified by the developers in an earlier stage. While this method allows any latent categories to emerge, its empirical, highly technical nature is daunting to most action researchers. Clearly a procedure that utilizes the strengths of each model and provides a method for teachers-as-researchers to establish content validity evidence is required. This paper utilizes a variation of the two procedures in which judges are provided with items on separate cards and asked to sort the cards into meaningful categories. However, a simpler analysis of the responses is utilized to determine relationships among the items.

Method

Data source. Two different groups of graduate students who are enrolled in a graduate program leading to certification as a reading specialist acted as expert reviewers for the content validation stage in the development of the Reader Self Perception Scale (RSPS). The first group of graduate students (n=30) reviewed the items using the forced-choice judgmental process described by Gable and Wolf (1993). The second group (n=33) reviewed the items using a latent category judgmental review procedure modified from Wiley (1967) and sorted the items into whatever meaningful categories they "saw" in the items. In addition, the RSPS was administered to 2,733 fourth, fifth and sixth graders.

Instruments. The RSPS is a recently developed scale that measures how children feel about themselves as readers (Henk & Melnick, 1995). Children respond to each of 33 items representing their perceptions of (1) their own *progress*, (2) *observational comparisons* they make relative to others in the class, (3) *social feedback* they receive from their peers, teacher(s), and family, and (4) their *physiological state*--that is, how they feel "inside" when asked to read. Strong alpha reliabilities ranging from .81 to .84 indicate a high level of internal consistency reliability in the instrument (see Table 1).

Procedures. The first group of graduate students were given the conceptual definitions for each of the four scales represented on the Reader Self Perception Scale (RSPS). They were asked sort each of the 33 items into the category it seemed to fit best and to indicate how strongly they felt about placing the item in that category. Reviewers were provided with a fifth category called "Other" and instructed to assign any item that did not fit the first four categories into this one. The data were analyzed according to the procedure outlined by Gable and Wolf (1993).

A second group of graduate students were each given a deck of 33 cards with each card containing one item. They were asked to sort the cards into whatever meaningful categories they thought appropriate and, after final sorting, to describe the conceptual definition of what they believed each of their categories represented. Because each reviewer may have matched different combinations of items with each other, the proportion of reviewers who matched pairs of items was examined. All possible pairs were utilized that had at least 70% agreement.

The content validity results, (forced-choice and latent category methods) were compared with an analysis of scale reliabilities (Cronbach's Alpha) utilizing data from the RSPS which was administered to 2,733 fourth, fifth and sixth grade students.

Results

Table 1 presents the reliability results for each scale. The scale reliabilities were .81 for Social Feedback, .82 for Observational Comparisons, and .84 for both the Progress and Physiological States scales. As can be seen in the third column (Alpha if Item Deleted), all but two items contribute to the overall scale reliabilities. Item 10 in the Progress scale has a modest inter-item correlation and the alpha would increase slightly if the item were deleted. Item 5 on the Physiological States scales has a somewhat low inter-item correlation and the alpha would increase by 3 points if the item were deleted.

Although all items were placed in the appropriate *a priori* categories by 90% or more of the forced-choice content reviewers, the results of the latent category review yields slightly different results. Tables 2 through 5 contain the percent of agreement by content reviewers for all possible pairs of items. A criterion level of 70% agreement was established before a pair of items could be

included in the matrix. As can be seen in Tables 2 through 4, reviewers saw strong relationships among the items of the Observational Comparison, Physiological States, and Progress scales. Each of the three matrices for these scales indicate a high percentage of reviewers associated the items with each other. However, the Social Feedback scale matrix (Table 5) yields some interesting combinations of items. The latent category reviewers distinguished these items in three subsets--feedback from teachers (2, 3, 17), feedback from family (7, 9, 30) and feedback from peers (12, 31, 33). Even though the reliability analysis suggests that all items are inter-correlated sufficiently and contribute to the overall scale reliability, such sorting by expert reviewers may suggest that the content of the Social Feedback scale may indeed need to be further partitioned into those three sub-categories.

Conclusions

A comparison of the results of the forced-choice judgmental review and the latent category review provide an interesting contrast. While all items were placed in the anticipated *a priori* categories by the forced choice reviewers, latent category reviewers identified finer distinctions among the items. "Driving" the content review by providing reviewers with operational definitions may provide fewer distinctions among latent constructs. Although either method provides developers with a degree of content validity evidence, the latent category procedure may provide more accurate information.

Educational Implications

Content validation should receive the highest priority during the process of instrument

development. As the use of researcher-developed instruments by educational researchers increases, greater emphasis must be placed on appropriate methods to establish content validity. Procedures that take advantage of experts' content review insights can only strengthen the process and, ultimately, the instrument.

Table 1
Alpha Internal Consistency Reliabilities by Scale
(N=2,733)

Item Number	Item-Total Correlation	Alpha if Item Deleted	Scale Alpha
Progress			
10	.40	.85	
13	.54	.83	
15	.59	.82	
18	.69	.81	
19	.56	.82	.84
23	.64	.81	
24	.67	.81	
27	.43	.84	
28	.61	.82	
Observational Comparisons			
4	.62	.78	
6	.64	.78	
11	.68	.77	.82
14	.42	.82	
20	.69	.76	
22	.47	.82	
Social Feedback			
2	.45	.80	
3	.53	.79	
7	.50	.80	
9	.58	.79	.81
12	.51	.80	
17	.59	.79	
30	.51	.80	
31	.51	.80	
33	.48	.80	
Physiological States			
5	.31	.87	
8	.65	.81	
16	.71	.80	
21	.59	.82	.84
25	.70	.80	
26	.70	.80	
29	.52	.83	
32	.55	.82	

Table 2
Percent of Agreement for All Pair-Wise Comparisons
by Latent Category Expert Reviewers
Observational Comparison Scale
(N=33)

Items	4	6	11	14	20	22
4	--					
6	73	--				
11	82	85	--			
14	76	76	82	--		
20	88	79	91	79	--	
22	88	79	88	79	94	--

Table 3
Percent of Agreement for All Pair-Wise Comparisons
by Latent Category Expert Reviewers
Physiological States Scale
(N=33)

Items	5	8	16	21	25	26	29	32
5	--							
8	76	--						
16	73	91	--					
21	70	88	91	--				
25	70	88	91	91	--			
26	79	85	88	85	85	--		
29	70	85	88	85	85	82	--	
32	73	85	88	85	85	88	82	--

Table 4
Percent of Agreement for All Pair-Wise Comparisons
by Latent Category Expert Reviewers
Progress Scale
(N=33)

Items	10	13	15	18	19	23	24	27	28
10	--								
13	82	--							
15	88	85	--						
18	85	91	88	--					
19	79	76	85	79	--				
23	82	82	88	82	85	--			
24	88	82	91	88	88	91	--		
27	88	85	97	88	85	88	91	--	
28	85	82	94	85	88	91	94	94	--

Table 5
Percent of Agreement for All Pair-Wise Comparisons
by Latent Category Expert Reviewers
Social Feedback Scale
(N=33)

Items	2	3	7	9	12	17	30	31	33
2	--								
3	73	--							
7			--						
9			76	--					
12					--				
17	82	82				--			
30			82	76			--		
31					82			--	
33					88			79	--

References

Delcourt, M.A.B., & Kinzie, M.B. (1993). Computer technologies in teacher education: The measurement of attitudes and self-efficacy. Journal of Research and Development in Education, 27(1), 35-41.

Gable, R.K. & Wolf, M. (1993). Instrument development in the affective domain (2nd Ed.). Boston: Kluwer Academic Publishers.

Henk, W.A., & Melnick, S.A. (1995). The Reader Self-Perception Scale (RSPS): A new tool for measuring how children feel about themselves as readers. The Reading Teacher, 48(5), 2-14.

Swanson, J.L., Tokar, D.M., & Davis, L.E. (1994). Content and construct validity of the White Racial Identity Attitude Scale. Journal of Vocational Behavior, 44, 198-217.

Wiley, D.E. (1967). Latent partition analysis. Psychometrika, 32(2), 183-193.